Computing Basics

Contents

Chapter 1

Introduction

This resource contains a compilation of *basic* things everyone should know about computing, on various related topics. The intent is to provide a relatively concise guide to the essentials of computing which will make the most impact in improving your computing experience.

These "must know" basics are in documented in a wiki and are also compiled into an ebook (PDF and EPUB). Find the latest version of this document online at: https://github.com/brianhigh/computing-basics/wiki

1.1 Intended Audience

The target audience is "non-technical", smart people who know little or nothing about the technical details of computing, the Internet, security, etc. or how any of it actually "works".

We assume that you know how to use a computer and it's accessories, how to use a web browser and how to create basic office documents. We also assume you are a busy person, either a student, a working professional or otherwise occupied.

1.2 Just our 2¢

This work is a compilation of opinions, offered in the form of suggestions phrased like, "*you should*". Additional explanatory material has been added to provide context and to "fill in the blanks" by summarizing generally understood computing concepts. This material is provided by our contributors, based primarily on their own experience of providing computing "tech" support and accumulating computing knowledge over the years as "IT" people. This is essentially top of the head content, compiled into a wiki format, organized by general topics.

As we do not typically cite individual contributors or reference the specific sources of their contributions, we should be clear that the views expressed in this resource are essentially "just our our 2¢". While we may link certain terms to Wikipedia articles and other other online works, we do so only for your convenience. We hope that you take the time to look up terms you do not understand or are unclear about.

If you disagree with an opinion or take issue with something expressed as fact, please share your views through the "issues" link on our GitHub site. If you wish to have a more academic or scholarly treatment of this material, we encourage you to seek out textbooks, journal articles, and other works produced by trusted and proven authorities.

Think critically about any computing opinions or advice you read or receive, from any source. This information is provided for educational purposes only. Please see our license and disclaimer if you are unclear about this.

1.3 About the Authors

A community of computer support professionals, enthusiasts, and our readers have contributed to this resource. Originally compiled by members of the University of Washington's TechSupport email list subscribers, this work continues to be maintained through community effort. Please share your feedback with the project through our online repository. If we incorporate your contributions into the wiki, we will list you as a contributor in our README file.

1.4 Copyright, License and Disclaimer

Chapter 2

In a hurry? Top Tips

Backups!

No recent backups? Do it now. Otherwise, you gamble, worry, panic, then cry.

Be resourceful!

Search the 'net, read the docs, experiment (cautiously), and ask for help (as needed).

Think ahead!

Plan for change, allow for failure, budget for maintenance, and schedule replacement.

Pay attention!

Check your security and privacy settings. Be careful what you download and post online.

Get organized!

Store files in sensible hierarchy of folders with meaningful names. Keep it tidy.

Centralize!

Store shared files on a secure server and minimize passing copies around (by email).

Be efficient!

Use keyboard shortcuts to minimize point-n-click. Automate with scripts and plugins.

Make it snappy!

Speed up your computer by closing idle apps and installing more memory (RAM).

Keep it simple!

Don't go overboard on technology. Sometimes it's better just to talk in-person.

Chapter 3

Hardware

3.1 Computer

Computer hardware consists of the computer itself and any peripherals (the accessories used with it). The computer, whether it is a desktop or a laptop, contains the "brains" — the central processing unit (CPU) — and the components needed to feed it, such as a power supply, random access memory (RAM), data communications bus, and input/output (I/O) ports.

The CPU, an integrated circuit (also called a "computer chip"), is where the information is processed and where calculations are performed. The computer may have more than one CPU and each CPU may have more than one CPU core. Since each core is able to perform one task at a time (per clock cycle), having more cores allows your computer to perform tasks in parallel (at the same time). As circuitry gets smaller, CPU designers are able to fit more cores in a CPU. A recent advance called hyper-threading allows a single core to act as two logical cores. A computer may also have a graphics processing unit (GPU) to render graphics such as video animation and perform other delegated tasks.

3.1.1 Graphics Processing

In most computers, the graphics processing, video decoding, etc. are processed using integrated graphics. As the name implies, the graphics processor (GPU) is integrated (inside of) the CPU. For most desktop applications, this is more than sufficient. However, some applications, such as video conversion, can benefit from a dedicated graphics card.

Aside from graphics processing, the GPU can perform other non-video tasks. This is known as GPU Computing. Typical applications are large scale data comparison, computation, and pattern recognition. However, to take advantage of GPU Computing, you need an application that supports it, along with a compatible, dedicated graphics card.

3.1.2 CPU Cooling

CPUs and GPUs get hot and so are mated with a heat sink, and often a fan as well, all clamped together in one unit. The hotter the processor, the greater the need for cooling. Some CPUs offer frequency scaling which allows the processor to speed up as needed. As it speeds up, it heats up, and this may trigger the fan to increase speed (and noise) to keep up the cooling.

If your computer "revs-up" to make a loud vacuum-cleaner noise, this is why. This increasing of CPU and fan speed is in response to greater processing, which is caused by a program that is doing a lot of work, like rendering video, or getting stuck in a buggy processing loop. Sometimes the computer becomes unresponsive when the CPU load has become so high that all processing comes to a halt, the computer "locks up", and the only recourse is to physically

shut down the computer and restart it. If the cooling system is not working (i.e., the fan is broken or removed) and the temperature of the CPU exceeds a certain limit, then the computer may shut itself down or reboot.

3.2 Peripherals

Peripherals include the devices external to the computer which generally feed information into the computer or deliver it from the computer. Examples are keyboards, pointing devices (like "mice"), monitors (screens), speakers, etc. Many peripherals connect to the computer using a Universal Serial Bus (USB) connector.

3.2.1 USB

USB ports can vary in speed, so if speed matters, check your computer's specifications to make sure your ports meet your needs. For example, if you have a USB 3.0 device, you would want to plug it into a USB 3.0 port on your computer, usually colored blue (matching the plug). While a USB 2.0 port (colored black or white) will work, you will not get the extra speed advantage of USB 3.0 if you use a USB 2.0 port.

3.2.2 Video

Video connections also vary. While the connector shape will clearly indicate which port to use, you may need an adapter to connect the computer with the screen or projector. There are a lot of different types of video connectors in use, but basically they come down to three of the most popular: VGA, DVI, and HDMI. Apple devices often use miniature video connectors, which means you need to carry around your adapters with you if you plan to connect to a device with a different connector.

3.2.3 Other High-speed Data Connections

In addition to USB, there are a number of less common peripheral connections for high-speed data transfer, such as Firewire (IEE1394), Thunderbolt, and eSATA. Firewire is most commonly used with higher-end digital video cameras, and other digital video interfaces. Firewire connectors are also commonly found on Macs. Thunderbolt is mostly seen on Apple systems, and is used as a high speed interface for external storage, displays, and more. As for eSATA, it's exclusively used for storage — for connecting to single disk or multi-disk arrays.

3.3 Memory and Storage

People often become confused when discussing computer storage and memory. The confusion arises because both of these terms are used for two very different components. Becoming clear on these terms will help make a huge difference in your ability to maximize your use of your computer.

3.3.1 Memory (RAM)

A computer's "main memory" is the temporary "short term memory" also called volatile memory. This memory will only store information so long as the power stays on. Once you shut off the computer, any information which was in the volatile memory is lost. Memory is used to temporarily cache data and applications that the computer is currently using or has "open". When you close programs and files you are done using, memory is released and becomes available for other uses.

So, memory really isn't *storage* in the common sense, as whatever is there won't stay there very long. Instead people usually just call it *memory*. The most common type of main, volatile memory is random access memory (RAM). RAM is installed in your computer packaged as integrated circuits, often on a memory module containing many of them.

3.3.2 Storage (Folders)

A slower, but more permanent type of "long term memory" is called secondary storage. This is what people more often refer to simply as *storage*. Files are stored in "folders", or more generally, just "storage".

The folders are logical containers which are physically implemented in storage devices. These may be connected to your local computer or may reside in a server on the network. Network storage containers may also be referred to as "shares", "volumes", "exports", or "network drives".

Examples of physical storage devices are hard disk drives and solid state drives. These devices are usually installed within the case of the computer and are meant to store information even when the computer is powered off.

Some disk drives are used as external devices for *expanded storage*. Other examples of off-line storage are flash memory "sticks" and optical disc drives, since these are often inserted only when needed.

3.4 Avoid Confusion

- ***Memory*** = RAM = *short term* = faster = "working memory" = *close* stuff to free it up
- ***Storage*** = folders = *long term* = slower = "file storage" = *delete* files to free it up

When your computer gives you an error message say, "Out of memory", the computer probably means that it has filled your *volatile memory*, otherwise known as *RAM*. When this occurs, the entire computer will become very slow and may "freeze up" ("lock up") altogether. This may happen if you load too much data into a statistical package, spreadsheet program, or graphics application. It may also happen if your software has bugs which result in memory leaks — errors which can consume inordinate amounts of memory without freeing it.

You can free memory by closing applications, files, and "browser tabs" which don't need to be open right now. The more you close, the more memory you make available. Releasing memory can make a huge impact in computer performance.

Alternatively, if you receive a "disk full" error, this means that your storage (disk) is full — you have exceeded its capacity or your storage quota. Sometimes your available storage space fills when an application is buggy and writes way more data to the disk than it should. More often, people will fill their hard drives with files over time, until eventually there is no more room left. When this happens, software can no longer write to the storage and will usually either halt with an error message or keep trying causing strange delays.

People usually free up storage space by removing old or temporary files no longer needed, or are backed up somewhere else. You may also use a software tool called a *disk cleaning utility* or similar to automate this task somewhat. There are also some tools that can take inventory of your storage consumption to tell you which folders consume the most space. This will help you quickly find the large files which you no longer need, so that you can delete or archive them to free up a lot of space with a minimum of fuss.

3.5 The Kitchen Analogy

A commonly used analogy to explain memory (RAM) and storage (disk) is the so-called "kitchen analogy". Imagine you are a cook (the CPU) cooking in a kitchen (computer). RAM is like a counter top where the food sits while you

are preparing it. It would be inefficient and expensive to leave food on the counter all the time, but it is certainly a handy place to keep the food while you work with it. A hard drive is like a pantry and refrigerator, where food stays in storage until it is needed, but is not as quick to access as the counter top. RAM is designed for fast data access, which can be expensive. The hard drive doesn't have to move data as fast, so it's cheaper. That's why hard drives have greater capacity than RAM, but are cheaper and why RAM is used as a temporary place to store data while being processed by the CPU.

3.6 Virtual Memory

You should see that the causes and remedies of memory and storage errors are entirely separate and help to clarify the differences between the two.

However, while we are now clear on the distinction between memory and storage, we have to mention the one feature, called virtual memory, which blurs the lines. This is just a file or partition of storage which is used as an "overflow" area for volatile memory.

If you fill up your memory, then the computer may start to "swap" data from memory to storage. This is very inefficient and therefore very slow. If you have a hard disk drive, you may even hear a lot of clicking, known as "thrashing" as the data is read from and written to the disk very heavily. The computer will usually become very slow when this happens. It is best to never need to use this virtual memory by making sure you have plenty of available volatile memory.

3.7 Hardware Upgrades

3.7.1 Memory

As prices drop and capabilities increase, people are more likely to replace a device than to upgrade it. Also, given the rise in popularity of laptops (and notebooks, network, ultrabooks, etc.), and inclusion of more "on board" components in desktop computers, devices are less upgradable on the whole. So, what is the easiest and cheapest way to improve a computer? Start with RAM. It is relatively inexpensive to "max out" the RAM in a device by filling all available "slots" with memory modules. This is usually quick and easy for the owner of the device to do, as the most manufacturers still provide easy access to the memory slots.

3.7.2 Storage

The other common upgrade to consider is to replace the hard drive with one having a greater storage capacity ("space") or a solid state drive (SSD) for faster performance. This is can be technically challenging for most people, not so much for the physical replacement, but for the work required to back up the old drive, and prepare the new drive for use. As you can imagine, this procedure also exposes your data to more risk. For these reasons, people are more likely to use an external drive if they need expanded storage capacity, or they may use an online storage service.

Chapter 4

Software

Software is the name for instructions for computing devices. Software is "soft" because the instructions are not physical entities like hardware devices. The instructions may be stored on physical media like a hard disk or USB thumbrive, just as a cooking recipe may be written on a piece of paper or printed in a book. However, the recipe itself is just an *idea* of how to perform a task. Likewise, a software program is essentially just a list of instructions (or a *model* that generates instructions) for the execution of a set of desired computing operations.

4.1 Application Software

As you use a computer, the software instructions that are executed on your behalf by the CPU, such as programs and apps, are called application software. Applications are the programs that serve a specific purpose for a computer user or are to be used for completing certain tasks, such as using the Internet, composing a text document, or working with data.

4.2 System Software

4.2.1 The Operating System

Applications run within a software environment called the operating system (OS). Examples include Microsoft Windows, OS X, iOS, Android and Linux. The operating system contains thousands of files, many of which are utility software or software libraries.

4.2.2 Kernel, Drivers, and Firmware

An operating system also has a kernel, which is the central software program that manages the data exchange between the CPU and the other components within a computer. The kernel communicates with those components using device drivers, which are small programs that provide a software interface to the hardware. Devices that contain integrated circuits of their own may store software in firmware that allows updates through a procedure called flashing.

Chapter 5

Programming

A programmer takes concepts about a how to perform a computing task and translates those ideas into statements. These are usually typed into a computer file using a text editor as plain-text source code. The statements are composed according to a certain language syntax and semantics. The rules for composing the program statements are defined by the programming language that is chosen for the task. Different languages are optimized for different uses. Some languages are more general-purpose than others.

5.1 Compiling Programs

Most programming languages use a syntax that is convenient for human programmers, but is not directly understandable (executable) by computers. To translate the program into executable program code, the programmer may use a compiler to create a new binary file which the computer can run directly. Any time the program needs to be run, that binary file can be read by the computer and executed.

Alternatively, the programmer may use an interpreter to perform both the compile and execute functions in a single step. This allows for greater convenience in development and can also make the program more *portable* as it can be run on any system for which an interpreter has been installed, without the need to compile the program for any specific computing platform.

5.2 Compiled versus Scripting Languages

Sometimes people categorize computer languages into compiled and interpreted (scripting) languages. However, please keep in mind that a compiler or interpreter could be written for any language. Such distinctions might be useful to describe how languages are commonly used, but should not reflect on the nature or quality of a computer language itself.

Some languages are implemented as a hybrid of the two approaches. Many popular "scripting" languages like Perl, Python, R, and others typically use some compiled routines (for performance) for those tasks which benefit from them, but otherwise will use interpreted code. In this way, programs written in these languages can offer the benefits of both performance and convenience. Java is typically compliled into a intermediate form called bytecode which is then executed in a virtual environment called the Java virtual machine (JVM).

Most system software and desktop applications are compiled, generally for performance reasons. Web application software and programs written by end-users (e.g., data analysis, systems administration) are often interpreted programs, generally for easier, more interactive, development. So, the way a language is used will generally determine whether the development tools for that language include compilers, interpreters, or both.

Chapter 6

Networking

6.1 Introduction

Computers, phones, tablets, etc. communicate using wired and/or wireless networks. The most famous network is known as the Internet. Network communications are facilitated by various protocols, for example HTTP ("web") and SMTP ("email"), Ethernet ("wired"), and WiFi ("wireless"). The networking technologies and their protocols are designed to be modular and are organized into several protocol layers.

The use of standards (technologies, protocols, frameworks, etc.) makes networking easier to use and troubleshoot. Knowing a little about these standards will help you make better use of computer networks. These technical standards are defined in public documents (RFCs, etc.) and are developed openly by international working groups. Anyone can read these documents to understand the protocols and learn how they work together.

6.2 Reliability

Computer networks are useful, but not entirely reliable. If it is really important that a communication is made, then email, text, etc., may not be sufficient. Use a telephone or some other means of real-time communication if reliability is important. Although using computers or the Internet may be a convenient way to communicate, it may not be the best way in certain circumstances.

6.3 Wired vs. Wireless

All else being equal, for a typical computer workstation or laptop, a wired (e.g, Ethernet) network connection will generally be more reliable, faster, and more secure than a wireless (e.g, WiFi) one. If you are using a wired connection, but your device also has wireless enabled, you will save energy and possibly improve your network performance by disabling the wireless while you are wired. Some newer devices will do this for you by default, but verify this behavior with your system to be sure.

6.4 Device Management Tips

Keep your devices up-to-date on security patches, plugins, firmware, anti-virus, etc. While this is a standard security practice, it will also help fix certain problems caused by version incompatibilities and bug fixes. Be careful, though, as

some bug fixes and updates may create bigger problems than they solve. Some firmware updates can brick a device. If in doubt, search online for bug reports about updates before applying them.

Manage your personal WiFi network(s) and know how to log in to your router's administrative features. Make sure to consult the router manual and configure the highest encryption available, and use a complex password featuring case sensitivity and special characters for both the router's WiFi network and the router's administrative features.

6.5 Troubleshooting Tips

6.5.1 General Concepts

Do network troubleshooting starting with the devices closest to your machine and then move outward. For example, try accessing your own router (via ping or its web interface) or another machine on your local network before concluding your Internet connection is down. That way if it's a local problem, you don't waste time on hold with your service provider's technical support.

Given that professionally managed networks are generally well designed and maintained, a poor network connection is most likely caused by a fault at your end (your device), especially if you are using a wireless connection. The most common exceptions would be a network outage at your service provider or somewhere else along the route, but those will be relatively rare compared to problems you may have with your own equipment.

Out of all of the equipment in the chain, your personal equipment is usually the weakest link. The reason is that most people are using consumer-grade equipment (desktops, laptops, phones, tablets, home routers, etc.) with relatively cheap components, more unstable software, in a less protected environment, whereas the rest of the devices in the chain are generally industrial-grade telecom equipment and servers built with more rugged components, housed in more secure locations with adequate cooling and redundant power, and are more rigorously tested, monitored, and serviced.

All of the various links in the route between you and the remote site need to be working properly. If your computer seems to be fine and the site seems fine, but there is still a problem, the cause may be a network router or switch between you and the other system.

Take some time to learn about concepts such as domain name servers, routing, NAT, etc., as this knowledge will help your troubleshooting efforts immensely, saving a lot of time and frustration (yours and those who help you).

6.5.2 Troubleshooting Tools and Techniques

If you are having a problem with an application that uses networking, such as your web browser or your email client, then check networking with another application, website, device, etc. Try to narrow the possible causes. There are many layers of technology that must all be working together properly, and by limiting the variables, you are more likely to determine the most probable source of the problem.

First, if you using a wireless connection, make sure the wireless feature is enabled, switched on, etc., or if using a wired (Ethernet) connection, check that your network cable is plugged in (both ends), then check the link light on your computer and the network equipment (router, switch, etc.), making sure that any network devices are powered on. Finally, check to see if you have been assigned a valid IP address.

Know how to find your IP address, MAC address and DHCP settings. These are key pieces of information needed to get help from your network service provider. You can do this with utilities included with your operating system.

Learn to do a network "ping" and how to interpret the results. This is an invaluable troubleshooting tool that comes with most popular desktop and laptop computer systems. Ping and traceroute (tracert) can help you find the network

hops (routes) which may be down or slow. Again, these utilities come with popular computer systems. You will find many tutorials online regarding their use with a quick Internet search.

Just because you can make a "good" wireless connection, does not mean that the router (or access point) you have connected to will actually provide you with Internet access. Its own connection to "the outside" may be down, it may require some sort of authentication (or payment), or it may simply not be configured properly.

6.5.3 Performance

If your wireless connection is slow, check on the quality of the connection (how many "bars") and also consider interference (from microwave ovens, other network users such as in a large lecture hall, etc.). If possible, move closer to the router (access point) with a more direct "line of sight".

Even if the network is working perfectly, you still may not get the results you are expecting. You may see slow data transfer on what should be a fast network connection. The bottleneck may be the application you are using, extra computational overhead such as compression or encryption, network congestion on one or more links along the route, or simply that there are a lot of hops between you and the destination, each one adding additional overhead and latency. Just because the endpoints have fast connections does not mean that all of the links between them also have fast connections. As an example, major universities usually have very fast Internet connections, but a route between universities may have to use slower links to make the end-to-end circuit complete.

If your browser is slow to open or load the start page, you may want to choose a "blank page" or a simpler page as your start page. Likewise, web pages with large images, animation, advertisements, or interactive content will often be slower to load than simple pages with basic content. You can use browser plugins such as FlashBlock, AdBlock, and NoScript to limit the amount of extraneous content that your browser will process as the page loads. This will speed up your browsing experience and also reduce the load on your network connection and your computer.

Chapter 7

The Internet

We mention the Internet elsewhere in this resource, such as in the security and networking topics. But what *is* the Internet and *how* does it *work*?

7.1 The Web Analogy

The Internet is like a spider "web" with lots of paths to get from "A" to "B". This is a very well-known analogy, and inspired the name of the WorldWideWeb, commonly known as "The Web". However, make no mistake; "The Web" is *not* the same thing as "The Internet"!

Figure 7.1: Photo: Abhijitsmiles, CC BY 3.0 - Spider web with dew drops

7.2 The Web

Though the Internet had been around for many years before the invention of the WorldWideWeb, many people didn't know about it, so their introduction to the Internet was through the web and web browsers. Naturally, many people assumed the two were synonymous, and this led to some confusion. Whereas the Internet is a network infrastructure of communications links and nodes, the WorldWideWeb is a network of information and applications connected by and hosted upon the Internet. Web browsing is just one of many ways to use the Internet.

7.2.1 From Hyperlinks to Apps

The web began as a distributed collection of documents, hyperlinked through tags embedded in the documents. The tags are codes in a markup language called HTML. Originally, the only interactivity was to click on a linked word to load the page that the word linked to. Later, further interaction was introduced with scripting. Now, we have very interactive network applications, or web apps, of all kinds running in our web browsers. The benefit of using web applications is that you do not need to install anything other than your web browser (and maybe a browser plug-in) in order to use them.

7.2.2 Web Technologies

Web applications make use of many technologies, some of which are built into the Internet infrastructure. These network technologies carry the bits of information around the internet from server, to router, to your computer. Layered on top of these lower-level technologies are application-level ones, such as the various languages (HTML, CSS, Javascript) of the web pages. These are the languages found in the source code of each web page.

7.2.3 Dynamic Web Pages

Additional languages run on web servers to create web pages as you need them. These dynamic pages, created "on the fly" to suit your individual requests, are generated by server-side code written in other computer languages (PHP, JSP, ASP, etc.). Further, your web browser may be able to communicate with the server by passing data "behind the scenes" without needing to load a new page. This is commonly done with languages like XML and JSON.

7.2.4 Web Applications

All of this technology allows a web page to respond to your mouse movements and keyboard entries. As web applications become more responsive through techniques like Ajax, they become more like traditional desktop applications. A site can go "live" and offer a web app immediately, without the need for people to purchase, download, and install any additional software. This is extremely powerful, as it removes many barriers, and enables web apps to gain popularity "overnight". This has greatly accelerated social networking sites in particular, such as LinkedIn, Facebook, Twitter, Instagram, and Tumblr.

7.3 Internet Routing

Let's return to our earlier analogy for the Internet, of a network of paths connecting "A" to "B". Where the paths cross is analogous to an Internet router. Connecting the routers are the wires, fiber, and satellite links. The information is sent through the Internet in tiny data packets, traveling along these paths like postcards through the postal system.

7.4 Internet Map

7.4.1 Peer1 Map of the Internet

There are several ways to try and visualize the Internet. Pictured below are two views of how the Internet looks from an app (from Peer1.com).

> [...] the app's timeline is rooted in real data that uses timeline visualization to display 22,961 autonomous system nodes joined by 50,519 connections based on Internet topology from our partner in this project, CAIDA.
>
> — Rajan Sodhi *PEER 1 Hosting Launches Map of the Internet App*

Both views, "Globe" and "Network", were captured from the screen of an Android smartphone after doing a search for "University of Washington" (UW). The dots are Internet sites and the lines are connections between the UW and some other sites.

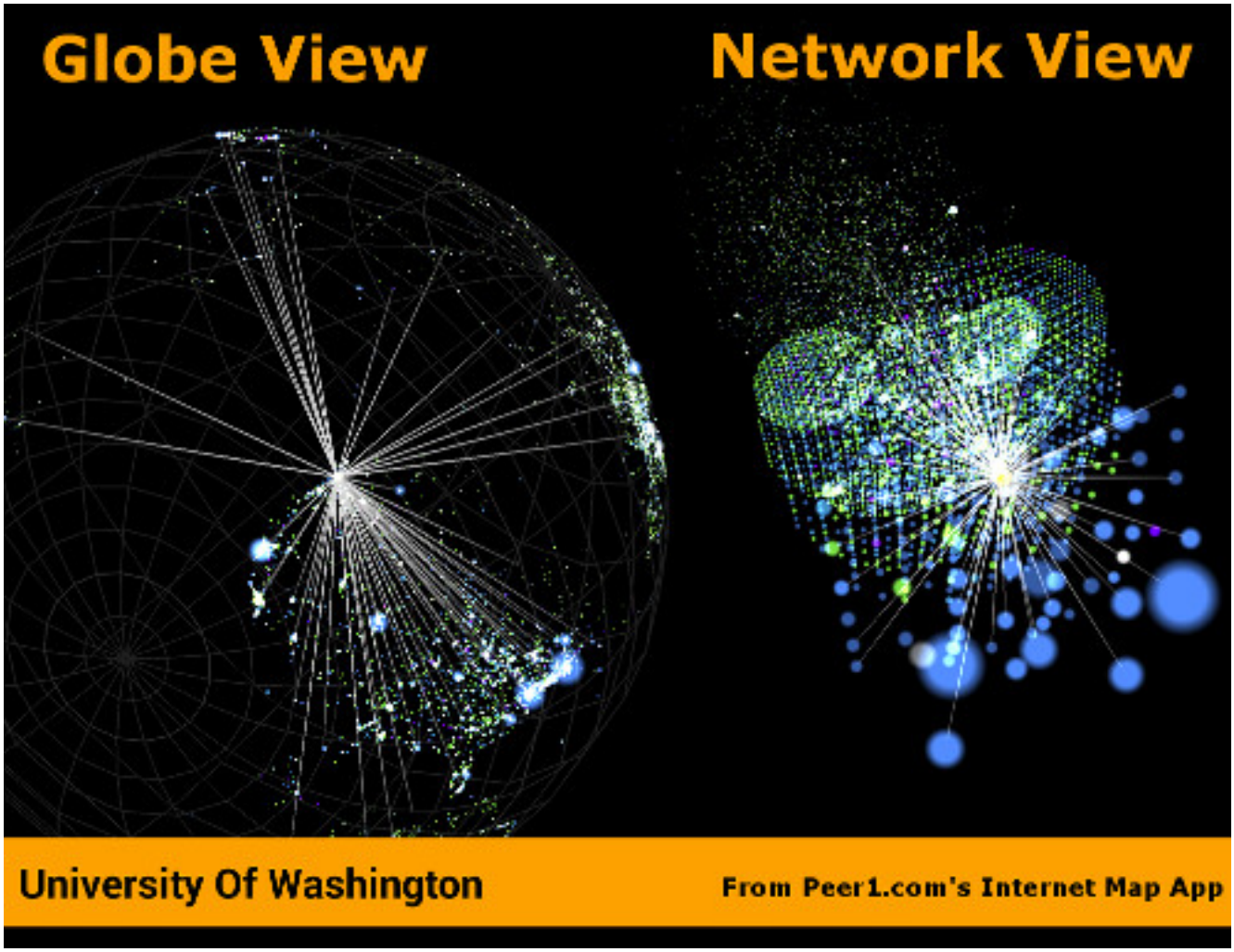

Figure 7.2: Image made with: Peer1 Map of the Internet app

7.4.2 internet-map.net

Another similar map can be found at The map of the Internet. As it is actually based on web traffic only, it should really be called *The map of the Web*.

> Every site is a circle on the map, and its size is determined by website traffic, the larger the amount of traffic, the bigger the circle. Users' switching between websites forms links, and the stronger the link, the closer the websites tend to arrange themselves to each other.
>
> — internet-map.net *The map of the Internet*

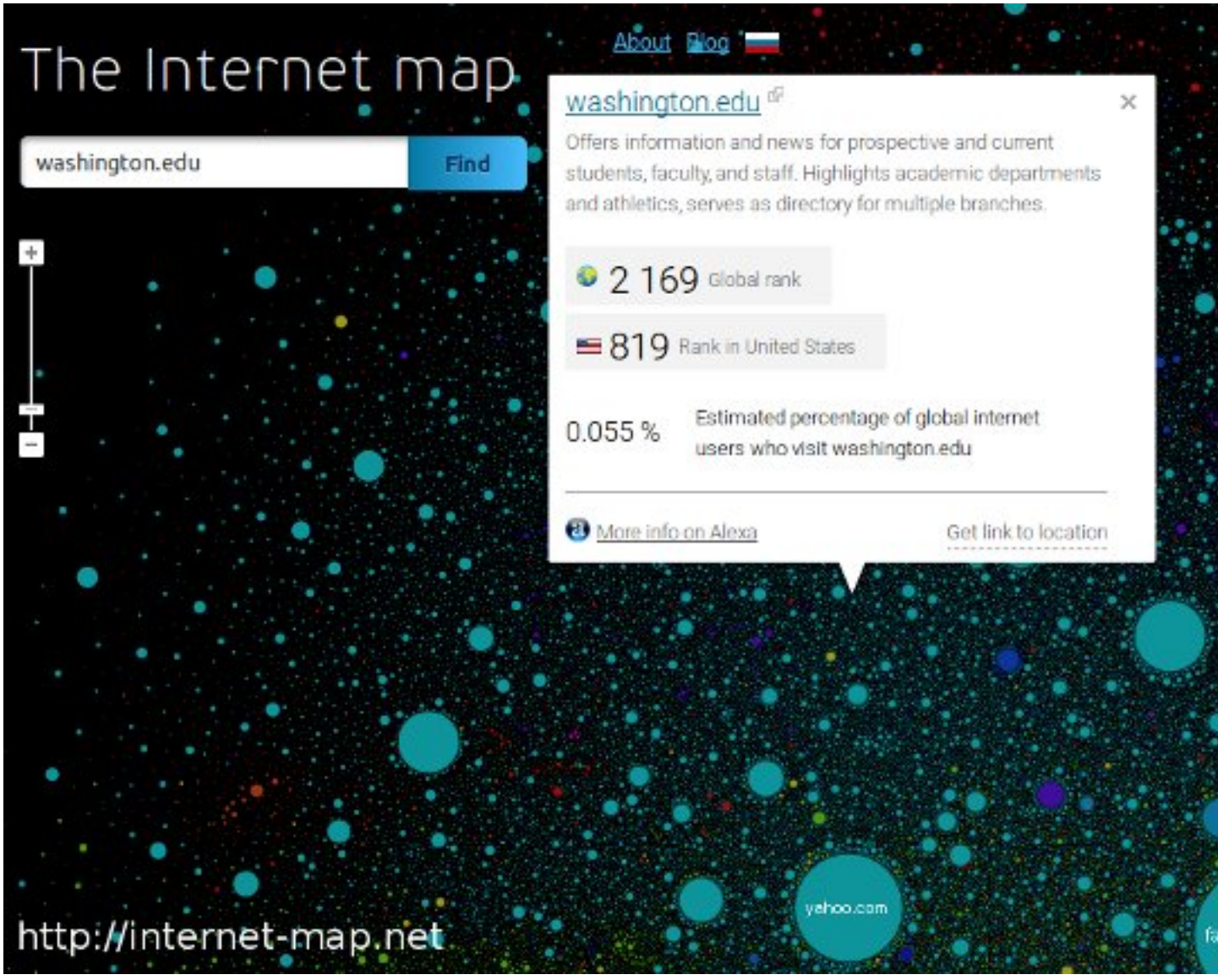

Figure 7.3: Image made with: internet-map.net

Chapter 8

Security

8.1 Introduction

Information security is about protecting information from unauthorized use as well ensuring availability for authorized use. *Computer security* is this practice applied specifically to computing devices, networks, services, and data.

As computer security is a "practice", not a "product", it depends on people, policies, training and behavior every bit as much as (and arguably much more than) software and hardware.

8.2 Be Smart

Some believe that security and convenience are mutually exclusive, that as one increases, the other will decrease. The presumption is that security measures make a system harder to use. This is not necessarily true, but having some degree of both security and convenience requires a smarter approach, carefully engineering the system so that the most secure behavior is also the easiest and most compelling.

Encourage a security-minded culture in your home or workplace. This will make it easier to develop and practice secure habits.

You are ultimately responsible for the security of your computing experience. Take an active role. Learn and understand basic security concepts. Engage in the computing behaviors or "hygiene" which will give you the level of security you need while still maintaining the degree of usability you desire. Be smart.

8.3 Manage Risks

Like the old adage, "out of sight, out of mind", risks not seen are easier to ignore. "Ignorance is bliss" . . . *until it isn't*.

Make an effort to see the practical risks of various online behaviors and how they might put you at risk personally, financially, and socially.

8.3.1 Personal Risk Assessment

Ask yourself, "What do I have which I need to protect? How valuable are those things to me?"

Consider the value of your property, your personal safety, your legal identity, your social reputation, your financial assets, your employment status, and your career/future. All of those, and those of your friends, family members, coworkers, employer, etc. are all valuable assets which you may put at risk with your online activity.

Consider threats such as identify theft, compromised bank accounts, stolen credit card numbers, stolen account credentials, investment scams, computer virus infections, loss of data, internet stalking, and disclosure of personal information resulting in social embarrassment, discrimination, persecution, hate crimes, loss of employment, property theft, or personal injury.

Evaluate how vulnerable you might be to each of those threats. This depends largely on your behavior. You can estimate the risks you face as the product of Threat $\times$ Vulnerability $\times$ Asset Value.

Now modify your behavior (including computing practices and online habits) to manage those risks.

8.4 Tools and Techniques

8.4.1 "End-users" View

When most computer-savvy people think of computer security, they think of: usernames, passwords, anti-virus software, security patches, firewalls, and encryption.

8.4.2 "Computer Administrators" View

Computer administrators and other computer professionals will also think of configuration: disabling unneeded services and accounts, changing default passwords, tightening access controls and firewall rules, strengthening security policy settings, alerts, logging, monitoring, backups, redundancy, and load balancing.

They also take physical security measures such as installing locks, cameras, and alarms. They often have to prove their systems are compliant with various regulations, so they will also think of documentation, audits and penetration tests. Further, they will stay current with the latest security news alerts about new threats and respond accordingly.

Regardless, all of these tools and measures are useless if people subvert them with insecure practices. So professionals will also create policies, find ways to enforce them, and educate their organization about the importance of secure practices. In this way, they encourage a culture of secure computing.

8.4.3 Your View

Since the practice of security involves addressing risks in all links of the chain, you do not want to be the weakest link. There is no reason why you should limit your practices to only those commonly known by "the masses". Consider investigating and utilizing the practices, tools and techniques of the professionals.

8.5 Best Practices

The majority of the "best practices" listed below came directly from our contributors, based on their professional and personal experiences, as well as their general knowledge of the practices commonly promoted by information security professionals. The University of Washington's Smart Computing page was also consulted as a reference. You are encouraged to compare this list against the many available online.

8.5.1 Basic Security Essentials

- Secure devices with locks, passwords, encryption, anti-virus software and host firewalls.

8.5.2 Software Installation and Updates

- Keep software updated, especially anti-virus software.
- Maintain your computer just like you do your car or yourself. If you neglect any of these, they will eventually fail.
- Exercise caution when installing *anything*, especially *free* or *shareware* software hosted by popular, often glitzy download sites.
- Some software installers come bundled with stuff you don't want so it's best to actually look at and read the prompts rather than just clicking *Next*, *Next*, *Next*.
- If you are prompted to update software, make sure it is a legitimate prompt before you agree to the update. Be wary of fake updaters for plugins, like bogus "Flash" updaters which may install malware.

8.5.3 Accounts and Passwords

- Log out or lock your screen when unattended. Otherwise someone could walk up and impersonate you - or worse.
- Don't share computer accounts. Make sure everyone has their own accounts. If you have shared your password with someone, change the password immediately.
- Use strong passwords. This means ones which are very long and/or very random. Mix upper and lower case letters, punctuation symbols and numbers. This increases the difficulty of cracking the password.
- Change a password promptly if it could have been seen by others, or if there is any indication that your account has been compromised.
- Changing passwords regularly may be required if password expiration policies are enforced.
- Don't use the same password for more than one account.
- If your passwords must be "written down", protect them with encryption in a password manager.
- Use password memory tricks to easily remember a different password for each site.
- Avoid telling your applications to "remember" your password.
- Don't use personal facts (such as birth date, birth place, etc) for answering security questions. A large number of personal facts are public record or readily available via social media.
- As an alternative, give answers that don't actually answer the question, e.g. if they ask for birth place, give them the color blue.
- Only use administrative accounts for specific administrative tasks. For ordinary, everyday activity, use a regular, non-privileged account. This limits the damage which can be done by mistake, mischief, or malware.

8.5.4 Data Security

- Know your data, safeguard it, and back it up regularly to multiple locations.
- Encrypt local backups if possible and practical, especially anything sensitive.
- Encrypt off-site backups, especially those stored on systems that you don't control, like "in the cloud".
- Purge data that you don't need to keep. Otherwise it is just a liability.

8.5.5 Internet Security

- Know that Internet-connected devices are always under attack. For example, see: Kaspersky Cyberthreat Real-time Map
- Check email links before clicking on them. Attachments and web links can hide attacks.
- If you receive a questionable attachment from someone you know, it's a good idea to contact them via a known contact method to confirm they sent it, before opening it.
- Check the encryption status of secure web pages with the "lock" icon (near the address bar in your web browser) and confirm that their web addresses begin with the "https://" prefix.
- Assume anything you post online can be read by anyone and can never be deleted.
- Configure the security and privacy settings on your social media accounts to limit exposure of personal information.
- Know that even with tight security and privacy settings you are still exposing yourself to some degree.
- Once a document (or any file) has been shared or sent, you've lost control of it.
- Before sharing something, be sure you can trust the recipient to honor any restrictions placed on it.

8.5.6 Security Culture, Training, and Help

- Take responsibility for your own security. Don't just expect IT, your ISP, or your vendor to do it for you.
- Take a basic computing security class, such as: NIH Information Security and Privacy Training Courses
- Don't assume that your family or coworkers practice safe computing. Helping them will help yourself.
- Your workplace may *require* you do actively take certain security measures and operate your computing equipment according to specific practices such as those listed here. Find out what those measures and practices are and put them into action.
- Regarding any suspicious computing event or security-related incident, if in doubt, check first with your IT support staff, or, if you have no IT support, then with the designated information security officer for your organization, if any, or with your supervisor or manager, or else with a *knowledgeable* coworker or friend, etc., your Internet service provider, software or computer vendor, or, at the very least, consider doing some investigation on the matter using reliable Internet sources.

8.6 Encryption

Encryption is the encoding and decoding of data. Usually this is done mathematically in software or in specialized hardware. It allows you to protect information so that some "secret" (key) is needed to access (decrypt) the data. While the encrypted data is safer when properly encrypted, one must carefully guard the secret key.

Encryption can be used for secure storage by encrypting files, folders, volumes, and disks. Encryption can encrypt data in transit, creating a protected tunnel that unencrypted ("clear text") data can pass through.

8.6.1 Network Encryption (in Transit)

Usually passwords are encrypted in transit when logging into a system. The system will also store a "password hash" on the system to compare against the password you have sent. The hashed password is one-way encrypted so that it can still be used to authenticate you, but cannot be used by someone else. That is, the administrator (or some attacker) cannot decrypt (look up) the password from the hash. If you forget the password, it must be reset.

Web servers can use encryption for data in transit. When you access a web page using an HTTPS address, the browser will attempt to set up an encrypted connection to the web server. If this fails, or there is a certificate error, you will see an error message. You may choose to ignore the error and proceed anyway. If you do, then most web browsers will indicate the status of the encrypted connection with a warning icon or discolored "padlock" icon by the web address in the address bar. Otherwise, if the connection proceeds without error, then there will usually be a padlock icon with some indication of success such as a "green" color or "closed" image. You may investigate the certificate status by clicking this icon and viewing the encryption details.

Many other network services use encryption. For any network site you visit or service you use, look into whether or not encryption is used and how you might configure it or validate that it is working.

8.6.2 File Encryption (in Storage, at Rest)

As for encrypting files, you can use the encryption features of archiving software like 7-Zip, WinZip, etc., use a more general-purpose tool like GnuPG, or the file or disk encryption features of your operating system or device.

People should encrypt sensitive files if stored locally and before sending them over the Internet or any other untrusted network. The person on the receiving end needs to know how to decrypt the files, so you will need to coordinate with that person to make sure that they have the software and decryption key and know how to use them. You need to find a secure way to send the decryption key. Public key cryptography systems like GnuPG solve this particular problem, but are a little more complicated to use.

So, with a simpler system where there is a shared key, you need to send the key using some other means where there is no reasonable chance that someone might intercept it. For lower-security needs, a telephone call might suffice, or meeting in person, or using a secure web server (e.g., with a one-time web link) as a means for key exchange.

Don't ever email passwords unless you use encryption like GnuPG to protect the password. However, if you have this type of encryption set up with the recipient, then you could just use GnuPG to encrypt the file and dispense with the need to encrypt a password. That is the whole point of public key cryptography systems like GnuPG.

8.6.3 What Encryption Isn't

Encryption is not some "magic pixie dust" that you can sprinkle on yourself to make you safe. You have to use it intelligently along with all of the other recommended security practices. Don't just rely on one tool like encryption to solve all of your problems.

8.7 Insecurity

Nothing is completely secure. You have to determine what you are trying to protect and what you need to protect it from, then manage that risk in a practical way that you can afford.

A computer with an Internet connection is constantly under attack. Automated "bots" are constantly scanning all internet addresses, including the one your computer is using, to find open "ports", or network services.

These attackers are hoping that your system has a buggy or misconfigured service that can be exploited to take control of your system and use it for finding more vulnerable systems, sending spam, phishing messages, malware, harvesting passwords, installing trojan back-doors, etc.

8.7.1 System Compromises

Assume your system will be compromised and that your data will be accessed without authorization. With this attitude, you will be motivated to take realistic measures to protect your systems and data instead of simply relying on faith in some product or feature — or just worrying without actually doing something about it.

This is where getting serious about backups comes in, especially encrypted, off-site backups allowing you to perform a "bare metal restore" — reinstalling the entire system from a backup onto a new, fresh, blank disk.

You cannot reasonably expect to "remove" or "clean" a virus, trojan, or other malware since you don't know what else might have been installed once the system was compromised.

Even if the system appears to be working normally again, that does not prove it is secure. There could still be a "back door" or "password sniffer", "key logger", etc.

Therefore, the best and arguably *only* course of action after a compromise (and after any forensic measures have been taken) is to completely "redo" the system from scratch with a fresh install of all all software.

Forensic measures include any investigation of the system you might do in order to find out what really happened or who might have compromised the system. Usually this means removing any storage device, like a hard drive, immediately and making a copy of it for analysis, leaving the original drive unmodified and disconnected during analysis.

8.7.2 Email Insecurity

Email is not private or secure. Even if your connection to the mail server is encrypted (scrambled) in transit, the email itself (in storage) is not, unless you took some very specific steps to encrypt the contents of the email.

Very few people actually encrypt the contents of their email messages, know how to, know that it's possible, or even know what it means. Practically speaking, this is the realm of geeks, hackers, criminals, spies, and the military. But some ordinary people do occasionally encrypt attachments, such as Zip files and PDFs. Generally, most people don't.

In this discussion, we will assume that your email messages are not encrypted (in storage). Even if they were, they would eventually be decrypted by the recipient, and you cannot control what happens to the message once that has happened.

Email messages are usually passed from server to server and router to router without any encryption (in transit). Even if the servers did use encryption to pass email, the message would be stored on the servers unencrypted.

Anyone with administrative access to the mail server, or has "hacked into" that server, could read your message. Most of the efforts to secure email are spent on preventing spammers from abusing mail servers, not on the privacy of your email messages.

Once the message arrives at its destination, in the mailbox of the intended recipient, on their computer, it can be read by anyone with administrative access to that computer. How much do you trust the recipient or the recipient's family members (or coworkers) to keep your "secrets"?

Therefore, think of an email like a "post card" and do not use email for sensitive communications.

Do not trust that the actual sender of a message is the address listed in the "From" header. Email messages are easily and often forged. So, as stated earlier, do not trust links or attachments in emails and confirm with the sender if in doubt about any email or attachment which you may have received from them.

While it is possible to digitally (cryptographically) sign a file or some text, as in an email, most people never do, know how to, etc., as with encryption. Here, again, a tool like GnuPG can be used, as well as other *public key infrastructure* (PKI) utilities. Many email clients have support for this built-in, or it can be added with a plugin. Using digital signatures can then be used to validate a sender and the contents of the message as being legitimate.

Do not confuse the use of *cryptographic* digital signatures with simply digitizing a written signature into an electronic file. The latter practice provides no assurances against forgery. A cryptographic signature, on the other hand can be verified using a *public key*. If a cryptographically signed document is modified after it is signed, then the digital signature will be broken. This can be detected by checking the signature. As you can see, digital signatures provide assurances against both forgery and modification.

Chapter 9

Resource Management

9.1 Introduction

Computing Resource Management is making the best use of computing resources such as CPU, memory, storage space, bandwidth, etc. From a IT project management perspective, this includes areas such as requirements analysis and capacity planning. The idea here is to predict resource requirements ahead of when you actually need them. You don't want to be scrambling for computing power at the last minute. And once the project is going, you want to know how to verify that you do have the resources you actually need and are using them most effectively.

9.2 A General Approach

Pick the right tool for the job.

Workstation or Server, define your goals and scope of the project. Take a hard look at the job at hand and evaluate what it will take to accomplish the task cost effectively.

Take into consideration:

- vendor longevity
- ease of use
- budget
- security
- user base
- performance
- maintenance

Start with software. Find a "tool", application or suite of applications that will meet your project goals without a lot of extras.

Pick an operating system that will run the tool of choice and provide a low total cost of ownership.

Last, pick hardware that will best host your OS and application of choice, will meet the current requirements, and have a high probability to meet the future requirements for the next five years.

9.3 Estimating CPU, RAM, and Storage Needs

9.3.1 Know Your Data

The first step in resource planning is to determine how much data you will be working with. How you determine this will depend on the source of the data. For data collected from an instrument, look at a single sample, and just multiply the amount of data collected by the number of samples you expect to collect. It's a good idea to include a margin of error, perhaps ten to twenty percent extra, just in case. If the data is coming from an outside source, the data provider may have a rough estimate, if they have provided similar data to someone else.

9.3.2 Know Your Tools

Once you know the amount of data, it's possible to work out some rough ideas about other resources needed. However, two more questions must answered. How quickly do you need your results? And, what tools do you plan to use?

If your tools require storing data in RAM for processing, then your RAM needs will be directly impacted by the size of your dataset. So, if you find yourself with tens to hundereds of gigabytes of data, it may be wise to reconsider the tools being used, as there may be alternatives that can process data in smaller chunks from disk. If you can't change software tools, then consider upgrading your hardware. For example, if your dataset is 64GB in size, you'll be needing 70GB+ RAM to cover data and overhead, assuming you must import all of your data into RAM at once.

9.3.3 Know Your Performance Requirements

How quickly you need your results will impact CPU and disk performance requirements. The faster you need your results, the faster the CPU and disk you'll want. But these requirements are also relative to your dataset size. A small amount of data, say 1GB, can be read fairly quickly from a standard hard drive, whereas 100GB of data will take considerably longer.

The other factor impacting CPU is the tool selected. A single-threaded tool for example, will benefit most from a high clock speed (GHz) CPU with just a few cores. Whereas a heavily multi-threaded tool, will benefit from many cores, of modest clock speed. To best determine the CPU needs of whatever software you use, you'll want to check the documentation or contact the software vendor.

9.4 An Example Scenario: Server Purchase

Imagine you are a researcher in a science or engineering department at a major research university. You want a new server for a new research study in order to perform data analysis. You look at some websites and see that the server prices are just within your budget. So, you go the IT dept. and ask them to make the purchase. Here are some of the questions your IT person might ask.

9.4.1 Software and Support Questions

- Will you need any software installed on that server, such as an operating system and data processing applications? What are those? Most with commercial licenses will require that you budget for this extra cost.
- How long will the system be in service? Can the warranty be extended to meet your requirements?
- You should also strongly consider a support contract for your hardware and software. This may cost roughly 10-30% per year. Are you willing to pay this or face being "unsupported" by the vendor?

- Will you need any custom software developed? Who will write this code and how much will it cost? Will the developer continue to support it for the duration of the research project? If the software is developed internally, what happens if the developer leaves the organization? Will they continue to support their code? Under what terms? Expect to pay roughly four times the initial development costs or more for code maintenance.
- Who will maintain (perform updates, upgrades, repairs, monitoring) the server?
- Who will be the primary contact person in the research group for ongoing support issues?

9.4.2 Performance and Reliability Questions

- What are the uptime/availability requirements?
- Will you need load balancing, clustering or other high-availability features?
- Will you need high speed network connections or data connections?

9.4.3 Hosting and Maintenance Questions

- What are the hosting requirements? Where will this server be hosted? How much will that cost?
- Power (UPS, generator, redundant circuits, conditioned power)? Cooling? Physical security?
- Will the server need to be expandable? Will you need more storage later? Long-term data archival?
- Who will maintain the server and perform backups? You or the hosting provider?
- What are the decommissioning requirements?
- What spare parts can you afford to purchase? Will those be purchased now, or will money be set aside for this?

9.4.4 Data Storage, Management and Backup Questions

- What kind of data storage requirements will you have?
- Will you need backups? Off-site? Encrypted?
- Have you budgeted for the cost of backups? Are you prepared to purchase backup hardware and software?
- How much data will need to be backed-up and how often? Will you need (incremental) snapshots?
- How long will backups need to be archived?
- What are the data management requirements? Does your research grant specify any?
- What are your data retention and data destruction requirements?

9.4.5 Security and Compliance Questions

- What are the security requirements for the project?
- Will you be storing personally-identifying (subject/patient) information? How will that be de-identified?
- Will more sensitive data need to be stored and accessed differently than less sensitive data? How will this be managed? Within an application or by the operating system?
- Will anyone (like software developers or database administrators) need special administrative access to the server? Will they also be certified to access any sensitive data?
- Will the server need to be compliant with any government standards such as HIPAA or FISMA? If so, are you prepared for the costs and delays involved in meeting compliance, including documentation and auditing?

9.4.6 Collaboration and Access Questions

- Will you need to share research data with others? What kind of access will they need? Who will manage that access (accounts, passwords, group memberships)?
- Will the server use local user accounts or will it tie into some centralized user account system within the organization? Will this account system include accounts for all collaborators, even those who come from outside of the organization? How will those people be able to access the server?
- Will all access be from campus or will some form of remote access (VPN, SSH, SFTP, Remote Desktop, etc.) be required?
- If collaborators need access to sensitive data, how will they be certified to access that data? How will the IT people know who is or is not certified when granting access to data?

9.4.7 Final Questions

- Are you prepared for all of these additional costs to equal or exceed the cost of the server itself? Have you budgeted for all of this? Is there enough money left? If not, then what?

Chapter 10

Files

10.1 Introduction

Your computing experience can be frustrating if files are not easily found, opened, or saved. We will explore file-related concepts, some common issues and offer some suggestions.

10.2 File and Folders

Files are stored on a computer in a nested structure, or "hierarchy", of folders (also called "directories") and subfolders. A subfolder is just a folder which is organized within (or "below") another ("parent") folder. Folders can be nested many levels deep.

The topmost folder is sometimes called the "root" of the folder hierarchy. A file placed at that top level is not considered to be in any particular folder. On some computer systems, such as Microsoft Windows, this top level may also be represented as a "drive", but may not actually correlate to a physical disk drive.

File structures may be stored on the local system hardware, on another system in a network, or distributed across many such systems. File structures are implemented in software as a "filesystem". Hardware devices like hard disks and flash memory devices are formatted with one or more filesystems before files are written to them.

10.3 File types and formats

10.3.1 Text Files

A computer file may be "text" or "binary". Text files are strings of characters from a standard character (such as ASCII).

Examples are:

- simple text (just characters)
- delineated text (characters separated with some special character, e.g. CSV, comma separated variables)
- structured text (like web page code, i.e. HTML or XML, or JSON)
- computer code (characters structured as a "program", i.e. source code)

The file name could be just about anything, but often they will end with a "suffix", usually a dot/period (.) followed with some characters representing the file type such as txt, csv, tsv, html, html, xml, c, py, pl, R, etc.

10.3.2 Binary Files

Instead of using just "plain text" characters, a file can also contain a mixture of characters or other non-character data, such as multimedia (images, video), compiled computer code (like an application executable), or compressed data of any type. They are called "binary" since their composition does not conform to any particular standard character coding system, and thus the file can seen as merely a string of binary digits (i.e., ones and zeros). Of course, to a computer, every bit of information appears binary, but this sort of terminology is meant for us humans.

Examples are (with example file suffixes):

- Compressed files (zip, tgz, rar, etc.)
- Multimedia files (gif, jpg/jpeg, png, mov, wmv, wma, mp3, mp4)
- Document files (doc, docx, xls, xlsx, ppt, pptx, pdf)
- Binary data files (Rdata, dta, mdb, sas7bdat, dbf)

Binary formats are sometimes defined in a formal standard, as is the case with many popular multimedia formats, while others are privately defined by vendors for use by their specific applications or products and are not well supported by other products.

10.3.3 Open versus Proprietary File Formats

If development of these formats is private and closed, the formats are often referred to as "proprietary". Alternatively, file format standards developed in an open, public, "community" context, may be called "open" or "open source". The same terminology is used for other technology standards such as for network protocols or software.

The advantage of open file formats is that they are more likely to be supported by a wider range of tools, applications, or products than a closed "proprietary" format. This makes files easier to import, export, and convert for use in alternative applications. While third-party developers may "reverse-engineer" closed formats to write applications which allow some degree of interoperability, they may not guarantee full compatibility.

10.4 Default Application

If you try to open a file by clicking on it, your computer will try to guess which application should be used to open it. Most computer operating systems will look at the filename suffix and compare that against its internal database of "associations".

Some operating systems, such a OSX, may also store an association for the file when it is created, regardless of the filename. Since not all operating systems operate this way, such as Windows, and as this association can be lost during file transfer, files shared with others should have a standard file suffix.

You can train your computer to use certain applications to open certain files or file types. This is called, "setting the default application". You can also simply open a file from within a particular application. This is a handy way to work around a broken or missing association. Some applications know how to open many different types of files.

10.5 Parsing and Converting

When files are read by an application, they are parsed in some way to bring the file's contents into your computer's memory, as an internal data structure. For the application to know how to parse the file, the file needs to be in a file format that it knows about.

If an application doesn't know about a particular file type, you will need to convert the format to a format it does know about. While you may change the suffix of the file by renaming it, this will not change the file format. To change the format, you need to convert it to the new format.

Some applications can open and export files with various formats. This is done using the familiar *File→Open* and *File→Save As* menu options, or similar.

10.6 File organization and naming

When storing files on your computer, it really pays to organize the files into a meaningful structure of folders and subfolders. What structure should you use? Assume another person needs to find your files and knows nothing about your folder structure. Create a folder hierarchy starting from general and going deeper into specifics. Try to avoid redundancy in file and path names. This will save you extra work frustration. If your work is project based, try something like this file path:

```
projects\{name}\data
```

The represents the levels between folders. So, {name} is an actual project name and "docs" is the folder for the documents relating to that project. If you work is organized by client, or by class, etc., then that should be at a higher level than the topics relating to those high-level divisions. For keeping track of coursework, you might try:

```
courses\{name}\assignments
```

Where {name} would be the actual course name, in this example file path.

As for the naming of individual files, preferences vary, but it is good idea to name the file with a succinct description of what the file is, as distinct from the other files in the folder. Names are a little restricted by allowed characters, those most are allowed these days. File length and path length are also an issue, but the limits are usually not an issue. You can find these limits by looking in the documentation for your operating system.

Here is an example of a problematic file path. Can you guess why?

```
misc\stuff\joe's files\joe's work files\temp\DON'T DELETE ME!\project 1\May\ ↩
   project 2\old\pics\joe.xlsx
```

10.7 File Sharing and Collaboration

Sharing files is common within organizations and among collaborators. Typically workgroups have access to a file server and sharing is simply a matter of working with the files as they are stored there. The idea is to work from a single copy of the file in a central location. This file server is backed up and so the individual users do not need to worry about this detail, nor do they have to pass copies around (e.g., through email). They still have to manage revisions to files which change.

Remote access to the file server may be offer via secure file transfer protocols such as SFTP, a virtual private network (VPN), or a virtual "desktop" session like Remote Desktop, VNC, or X2Go. Since file transfer tends to spread copies of the files around onto the computers of the various collaborators, the other remote access options are generally preferred for collaboration and security.

Many people are in the habit of saving a copy of a file when it is time to make a new version, keeping the original as the previous version. While this method works and provides a simple, but crude history of changes, there are other more sophisticated methods, such as "track changes" features and version control systems. Backups should not be

used as a versioning system, since system administrators usually use a backup rotation schedule which reuses backup media, replacing older backups with newer ones.

The central file server can also be a third-party "cloud" storage service, such as Dropbox, Google Drive, OneDrive, SharePoint, etc. These services offer a "free" tier and can be very handy due to "apps" for various devices, automatic synch, and design for mobile collaboration. The same concepts of server use mentioned above apply. However, off-site, third-party storage may not meet security, regulatory or service-level requirements.

The "cloud" services can offer compelling value over traditional file servers in that collaboration features (e.g., co-authoring, portals, workflows) are built-in, providing the experience of an integrated application, not just a file depot. Further, these platforms provide rights-management features in the form of "invites" which greatly facilitates user-controlled sharing.

10.8 File Corruption and Repair

With frequent backups, you should not have to wrestle with repairing corrupt files. Just restore them from backup. But if your backups are not sufficient, make a note to improve them and then try the following approach.

If a file appears to be corrupted, stop what you're doing. The corruption may be caused by a faulty drive or media. In which case, further activity may cause further damage.

1. Scan the drive or media with Windows Disk Check, or on a Mac use Disk Utility. It's also advisable to use a tool like Crystal Disk Info to see if the drive itself is reporting what are known as SMART errors.
2. If the drive is OK, then the file itself may have been corrupted by an application bug. Some applications "save" temporary copies of files in your computer's temp folder. In which case, it's wise to check it for a recent copy.
3. If there isn't a temporary copy, then try Windows Previous Versions (Shadow Copy), or OS X's Time Machine. If you're fortunate, they'll have been enabled, and may have saved a copy of the file. However, this option generally only applies to files on your computer's hard drive, not on portable media or network storage (which may have its own snapshot and backup systems).
4. If all else fails, you may have to try repairing the file. The repairing process depends on the format (software used) and version of the file. A search on Google for "repair X files utility", where X is the file type should reveal some options. Popular file types like MS Office usually have some kind of tool on the market.

Chapter 11

Data

11.1 Introduction

Data are the individual pieces of information we store in files and share through the network. So, what applies to files, such as the importance of backups, also applies to data. The same goes for security principles and practices.

11.2 Data, Documents, and Databases

We may treat data differently than other information. Whereas a document, such as a MS-Word file, may be intended only for human readers, raw data are usually meant to be read and processed by automated means — by machines. The data may be queried, analyzed, and summarized into tables and plots for human eyes, but most people do not want to see all of the raw data directly.

Data may need to grow immensely in size without slowing down its processing. This need for scalability requires data to be managed more carefully and thoughtfully than individual document files. This is why data are often stored in data structures called databases. Databases are specifically designed for efficient storage, searching, and processing of large amounts data. Sometimes, it is easier and more practical to store data in an individual data file than in a full-featured database system. It really depends on your needs.

11.3 "Up-front" Data Planning

Determine *up-front* how important the data will be and treat it accordingly. Take backups seriously and design and implement the best automated backup system you can afford, then regularly audit and test backups to make sure you can restore from them. Without good backups, you are one small mishap away from major disaster. Who wants to live like that?

Determine *up-front* how the data will be accessed. Strongly consider formatting your data for easy automated processing by using simple tabular structures of rows and columns in common, flexible file formats. Realize that you may want to collaborate with others and consider allowing for multiuser, simultaneous, and remote access. Sharing data files by email does not scale well and is insecure. A file sitting on a file server will not allow simultaneous editing, and one person's edited version will overwrite the whole file. Manually merging changes can be extremely difficult and error-prone. Linking separate files together properly can be challenging without specialized tools like relational databases.

Determine *up-front* how long the data will be around, and what the plans will be when it's no longer useful. If the plan is to create 500,000 files/week for a decade, do the math and figure out if it's practical to store and analyze these files before you're in year 8 and things start falling apart.

11.4 A Tidy Approach to Data Management

Start with an easily machine-readable file format (when possible), preferably an open and commonly-supported format like CSV. This will reduce the amount of work required to convert the data into a usable form for the widest number of data analysis applications.

Be consistent. For example, for every row of data containing dates, use the same date format. If a column contains a name or identifier, use the same capitalization or hyphenation rules throughout. Don't make the computer guess. This consistency will allow you to link related tables using a common (and consistently used) identifier. It will allow you to group data by categories so you can more easily calculate meaningful statistics and produce useful plots.

Normalize your data. Divide data into tables, each one containing data about a specific type of entity or observational unit. For example, if your study involves subjects, locations, samples, and test results, keep the information specific to each of those entity types in a separate table, and then link them on a common field, like an identifier (ID). This way, you are only repeating the ID in the linked tables. If you need to update the record for one location, for example, you will only have to update one record (row) in your location table. Each row should represent only one instance or observation, while each column should represent only one attribute, characteristic, measure, or variable. These are core principles of tidy data. Taking a little time to organize your data in this way will make working with it much less error-prone and far easier.

Use the right tool for the job. Often people are drawn toward software such as MS-Access and FileMaker due to the ease of use. However, such tools don't scale very well and perform poorly in multi-user situations. If you find yourself needing a relational database, it's best to start with something like MySQL or MS SQL, and use one of their many graphical database management tools. While they do have a somewhat higher learning curve, their ability to scale means you don't have to retool your work flow as your needs grow.

Similarly, if you are primarily doing data analysis in R, then consider using the data.table package to organize, search, and manipulate your data instead of using data.frames. You will gain many of the features of relational databases, all within your familiar R environment, without needing to learn yet another language or framework. Plus, you are likely to see increased performance and cleaner, simpler code.

So, if you're lucky, your favorite tool may support a *tidy data* approach, but be sure to allow for scalability, flexibility and collaboration. Don't be afraid to stretch a little and learn a new package or tool if it will help your long-term data management goals.

Chapter 12

Buzzwords

The technology world suffers from excessive use of buzzwords (words associated with hype, or "buzz"). Someone coins a term to describe a looming problem or seemingly magical solution. The term or phrase gains momentum with media exposure, and soon you see it everywhere.

Yet, often, people are unsure about what the term actually means or if it really matters. Relentless exposure to the ill-defined hype results in the modern malady known as *buzzword fatigue*. We will help clarify a few of them by peering through the clouds.

12.1 Cloud Computing

The notion of "the cloud" in computing comes from the use of a cloud symbol in network diagrams. The cloud symbol represents the "rest of the network" or the part of the network lying outside of your organization or your network diagram. Usually, this "outside" means the Internet. So, cloud computing is using servers outside of your organization, often over the Internet. As you may realize, this is not really that new, but more of a increasing trend.

You now have several competing services to choose from that present vast computing resources as a "commodity", like electricity. Computing resources may be rented on an as-needed basis to scale to meet varying demand. You only pay for the resources you use. You can even run your own virtual machine in the cloud, through the wonders of virtualization, a buzzword with its own family of buzzwords.

So, the attraction is not having to buy, house, and maintain hardware, networking, and in some cases, software. The potential downsides all relate to "lack of control".

- ***Security***: You have to trust the cloud service provider with your data.
- ***Support***: You have to go through the cloud service provider if something isn't working the way you expect it to work.
- ***Availability***: You have to depend on your network connection to the cloud service, in terms of both reliability and performance.

Some people consider "the cloud" as just another name for "the network" or "not my computer". In any case, the notion of *cloud computing* involves some degree of "outsourcing" of computing resources and accessing those resources through the network.

12.2 Big Data

Big Data is a term applied to very large amounts of data. Search online and you will find many different ways to define this. For our purposes, we will use the broad definition of: a collection of data so large that its largeness creates significant processing problems, but can yield value not found in smaller data sets.

Very large data sets are enabled by very low cost storage. For example, instead of sampling every hour or even every minute, you can sample every second because you have the space to store all of the resultant data.

12.3 Data Mining

Data mining is looking for patterns in a very large data set. Imagine trying to find the "needles in a haystack" of huge volumes of data. Much of this is enabled by very low cost, parallel computing.

The limiting factor here becomes the network because you can only move so much data in a given time frame. Here you may turn to *cloud computing* for massively parallel processing power.

12.4 Data Science

Data science is a term generally used to refer to statistical data analysis and the presentation of results, often visually. In particular, some people use "data science" to mean the analysis of *big data* using techniques such as *data mining* to produce stunning visualizations. The publishing of such *data products* is termed data journalism. Of course, the analysis may make use of *cloud computing* in terms of cloud storage and processing.